WEKA 3.5.5 用户指南

wizardforcel

Published with GitBook

目錄

介紹	0
1 启动 WEKA	1
2 WEKA Explorer	2
2.1 标签页	2.1
2.2 状态栏	2.2
2.3 Log 按钮	2.3
2.4 WEKA 状态图标	2.4
3 预处理	3
3.1 载入数据	3.1
3.2 当前关系	3.2
3.3 处理属性	3.3
3.4 使用筛选器	3.4
4 分类	4
4.1 选择分类器	4.1
4.2 测试选项	4.2
4.3 Class 属性	4.3
4.4 训练分类器	4.4
4.5 分类器输出文本	4.5
4.6 结果列表	4.6
5 聚类	5
5.1 选择聚类器（Clusterer）	5.1
5.2 聚类模式	5.2
5.3 忽略属性	5.3
5.4 学习聚类	5.4
6 关联规则	6
6.1 设定	6.1
6.2 学习关联规则	6.2
7 属性选择	7
7.1 搜索与评估	7.1
7.2 选项	7.2

7.3 执行选择 7.3

8 可视化 8

8.1 散点图矩阵 8.1

8.2 选择单独的二维散点图 8.2

8.3 选择实例 8.3

参考文献 9

WEKA 3.5.5 用户指南

原文版本 3.5.5

翻译：王娜

校对：C6H5NO2

Pentaho 中文讨论组 QQ 群：12635055

论坛：http://bbs.wekacn.org/

1 启动 WEKA

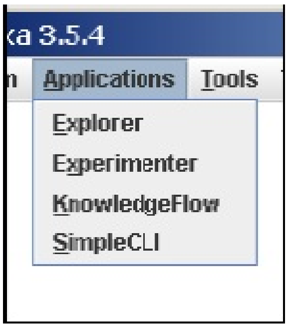
WEKA中新的菜单驱动的 GUI 继承了老的 GUI 选择器（类 weka.gui.GUIChooser） 的功能。它的MDI（ 多文档界面 ）外观，让所有打开的窗口更加明了。

这个菜单包括六个部分。

1. Program

- LogWindow 打开一 个日志 窗口， 记录 输出 到 stdout 或 stderr 的内容。在 MS Windows 那样的 环境中，WEKA 不是从一个终端启动， 这个就比较有 用。
- Exit 关闭WEKA。

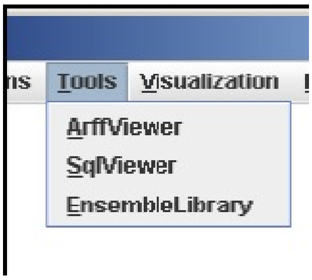
2. Applications 列出 WEKA 中主要的应用程序。



- Explorer 使用 WEKA 探索数据的环境。（本 文档的其它部分将详细介绍这个环境）
- Experimenter 运行算法试验、管理算法方案 之间的统计检验的环境。
- KnowledgeFlow 这个环境本质上和 Explorer 所支持的功能是一样的，但是它有一个可以拖放 的界面 。它 有一个 优势， 就是 支持 增量学习（incremental learning）。

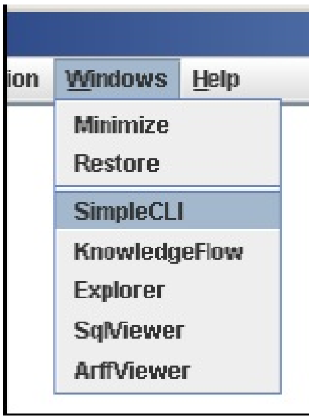
- SimpleCLI 提供了一个简单的命令行界面，从而可以在没有自带命令行的 操作系统中直接执行 WEKA 命令。

3. Tools 其他有用的应用程序。



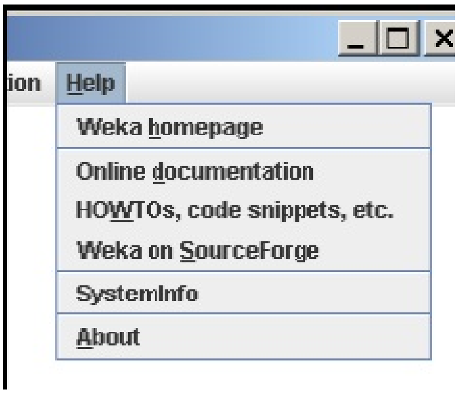
- ArffViewer 一个 MDI 应用程序，使用电子 表格的形式来查看 ARFF 文件。
- SqlViewer 一个 SQL 工作表， 用来通过JDBC 查询数据库。
- EnsembleLibrary 生 成 集 成 式 选 择（Ensemble Selection）[5] 所需设置的界面。

4. Visualization WEKA 中数据可视化的方法。



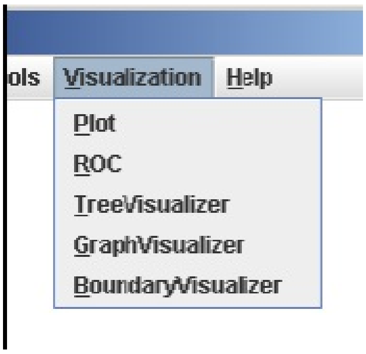
- Plot 作出数据集的二维散点图。
- ROC 显示预先保存的 ROC 曲线。
- TreeVisualizer 显示一个有向图，例如一个决策树。
- GraphVisualizer 显示 XML、BIF 或 DOT 格式的图片，例如贝叶斯网络（Bayesian network）。
- BoundaryVisualizer 允许在二维空间中对分类器的决策边界进行可视化。

5. Windows 所有已打开的窗口都列在这里。



- Minimize 最小化所有当前的窗口。
- Restore 还原所有最小化过的窗口。

6. Help WEKA 的在线资源可以从这里找到。



- Weka homepage 打开一个浏览器窗口，显示 WEKA 的主页。
- Online documentation 链接到 WekaDoc 维基文档 [4]。
- HOWTOs, code snippets, etc. 通用的 WekaWiki [3]，包括大量的例子， 以及开发和使用 WEKA 的基本知识（HOWTO）。
- Weka on Sourceforge WEKA 项目在 Sourceforge.net 的主页。
- SystemInfo 列出一些关于 Java/WEKA 环境的信息，例如 CLASSPATH。
- About 不光彩的 `About` 窗口。

如果从终端启动 WEKA，会有一些文字在终端窗口中出现。这些文字是可以忽略的， 除非某些东西出错了——这时它可以帮助找到错误的原因。（LogWindow 也可以显示那 些信息。）

这份文档也 可以从在线 的 WekaDoc Wiki [4] 中找到，它 将集中阐述 如何使用 Explorer，而不会逐个解释 WEKA 中的数据预处理工具和学习算法。要获得关于各种筛选 器（filter）和学习算法的更多信息，可参考 Data Mining [2] 一书。

2 WEKA Explorer

2.1 标签页

在窗口的顶部，标题栏下是一排标签。当 Explorer 首次启动时，只有第一个标签页是活动的；其他均是灰色的。这是因为在探索数据之前，必须先打开一个数据集(可能还要 对它进行预处理)。

所有的标签页如下所示：

1. Preprocess. 选择和修改要处理的数据。

2. Classify. 训练和测试关于分类或回归的学习方案。

3. Cluster. 从数据中学习聚类。

4. Associate. 从数据中学习关联规则。

5. Select attributes. 选择数据中最相关的属性。

6. Visualize. 查看数据的交互式二维图像。 这些标签被激活后，点击它们可以在不同的标签页面上进行切换，而每一个页面上可以执行对应的操作。不管位于哪个页面，窗口的底部区域(包括状态栏、log 按钮和 Weka 鸟) 仍然可见。

2.2 状态栏

状态（Status）栏出现在窗口的最底部。它显示一些信息让你知道正在做什么。例如，如果 Explorer 正忙于装载一个文件，状态栏就会有通知。

提示 — 在状态栏中的任意位置右击鼠标将会出现一个小菜单。这个菜单给了你两个选 项：

1. Memory Information. 在 log 栏中显示 WEKA 可用的内存量。

2. Run garbage collector. 强制运行 Java 垃圾回收器，搜索不再需要的内存空间 并将之释放，从而可为新任务分配更多的内存。注意即使不强制运行，垃圾回收 也是一直作为后台任务在运行的。

2.3 Log 按钮

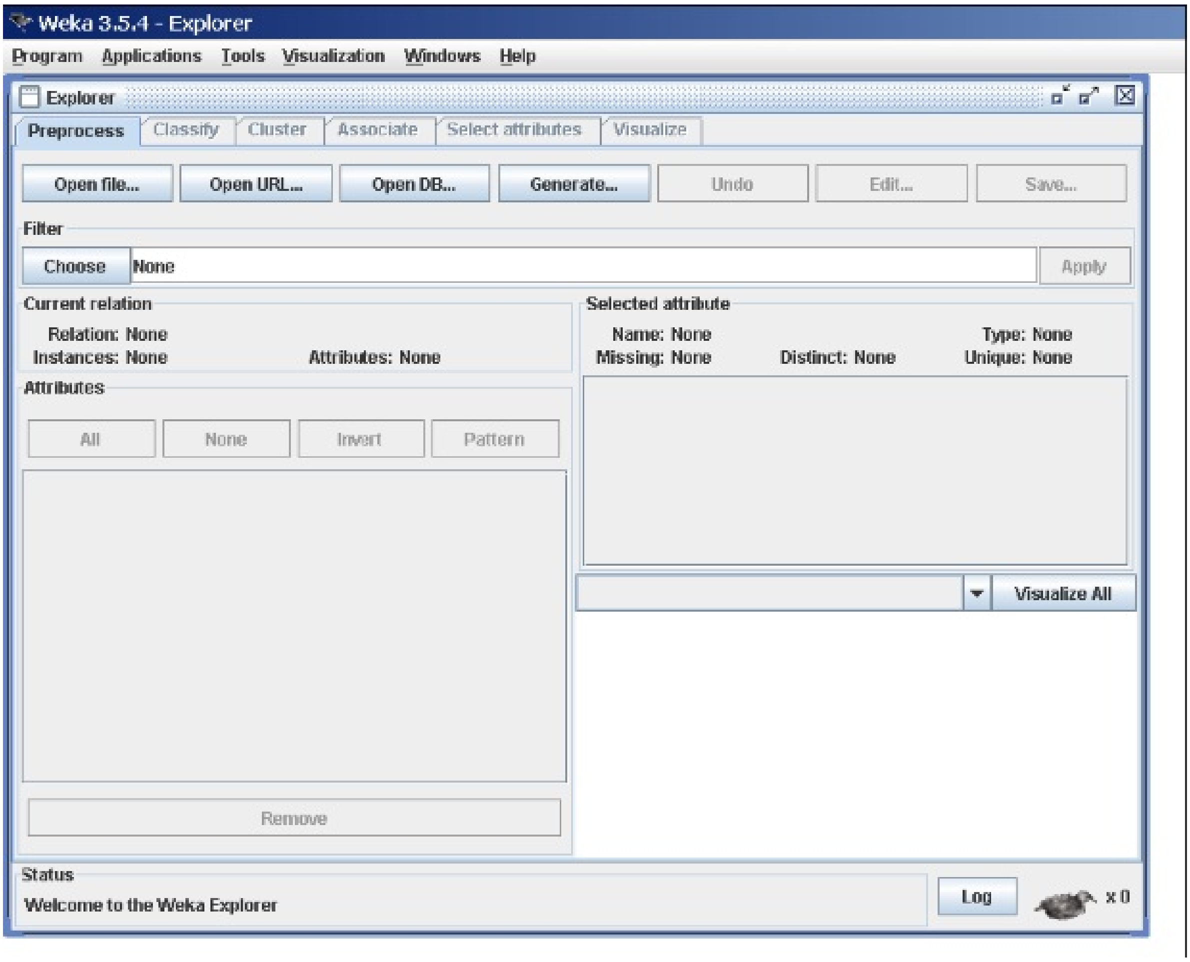
点击这个按钮，会出现一个单独的窗口，包含一个可拖动的文本区域。文本的每一行被加了一个时间戳，显示了它进入日志（log）的时间，一旦在WEKA 中执行某种操作时， 该日志就会记录发生了什么。对于使用命令行或者 SimpleCLI 的人，日志也将完整地记 录分类，聚类，特征提取等任务的设置字符，使得它们可被复制/粘贴到其它地方。但关 于数据集和 class 属性1的选项仍然要由用户给出（例如，分类器（classifier）的 –t，或 者筛选器的 –i 和 -o）

2.4 WEKA 状态图标

状态栏的右边是 WEKA 状态图标。当不运行任何进程时，WEKA鸟会坐下并打一个小盹。× 符号旁的数字显示了正运行的并发进程的数量。当系统空闲时，它是零，而当进程 的数量增长时，它也会增长。任意进程启动后，小鸟会站起来并到处活动。如果它仍然是 站着的，但是很长时间内不动，那么它生病了：某个地方出错了！在这种情况下，应该重 新启动 WEKA Explorer。

> 1 在分类或回归任务中，class 属性是默认的目标变量。注意这与下文中的分类型属性不是一个概念—译 注。

3 预处理


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Attributes
All None Invert Pattern
Remove
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Visualize All
Status
Welcome to the Weka Explorer
Log
x 0

3.1 载入数据

预处理页顶部的前4个按钮用来把数据载入WEKA：

1. Open file.... 打开一个对话框，允许你浏览本地文件系统上的数据文件。

2. Open URL.... 请求一个存有数据的 URL 地址。

3. Open DB.... 从数据 库 中读取 数据 (注意 ， 要 使之可 用， 可能需 要编 辑 weka/experiment/ DatabaseUtils.props 中的文件)

4. Generate.... 从一些数据生成器（DataGenerators）中生成人造数据。

使用 Open file... 按钮可以读取各种格式的文件：WEKA 的 ARFF 格式，CSV 格式， C4.5 格式，或者序列化的实例2格式。ARFF 文件通常扩展名是.arff，CSV 文件扩展名 是 .csv，C4.5 文件扩展名是 .data 和 .names ，序列化的实例对象扩展名为 .bsi。

3.2 当前关系

载入数据后，预处理面板就会显示各种信息。Current relation 一栏（`current relation` 指目前装载的数据，可理解为数据库术语中单独的关系表）有3个条目：

1. Relation. 关系的名称，在它装载自的文件中给出。使用筛选器（下文将详述） 将修改关系的名称。

2. Instances. 数据中的实例(或称数据点/记录) 的个数。

3. Attributes. 数据中的属性(或称特征) 的个数。

> 2 只有本段文字中的`实例`是 JAVA 语言中实例的概念；而后文中的`实例`都将指数据集中的记录— 译注。

3.3 处理属性

在 Current relation 一栏下是 Attributes（属性）栏。有四个按钮，其下是当前关系中的属性列表。该列表有3列：

1. No.. 一个数字，用来标识数据文件中指定的各属性的顺序。

2. 选择框. 允许勾选关系中呈现的各属性。

3. Name. 数据文件中声明的各属性的名称。

当点击属性列表中的不同行时，右边 Selected attribute 一栏的内容随之改变。这 一栏给出了列表中当前高亮显示的属性的一些描述：

1. Name. 属性的名称，和属性列表中给出的相同。

2. Type. 属性的类型，最常见的是分类型（Nominal）和数值型（Numeric）。

3. Missing. 数据中该属性缺失(或者未指定)的实例的数量(及百分比)。

4. Distinct. 数据中该属性包含的不同值的数目。

5. Unique. 唯一地拥有某值的实例的数目（及百分比），这些实例每个的取值都和 别的不一样。

在这些统计量的下面是一个列表，根据属性的不同类型，它显示了关于这个属性中储存 的值的更多信息。如果属性是分类型的，列表将包含该属性的每个可能值以及取那个值的 实例的数目。如果属性是数值型的，列表将给出四个统计量来描述数据取值的分布—最小 值、最大值、平均值和标准差。在这些统计量的下方，有一个彩色的直方图，根据直方图 上方一栏所选择的 class 属性来着色。(在点击时，该栏将显示一个可供选择的下拉列表。) 注意仅有分类型的 class 属性才会让直方图出现彩色。最后，若点击 Visualize All 按钮， 将在一个单独的窗口中显示数据集中所有属性的直方图。

回到属性列表，开始时所有的选择框都是没有被勾选的。可通过逐个点击来勾选/取消。 以上的4个按钮也可用于改变选择：

1. All. 所有选择框都被勾选。

2. None. 所有选择框被取消（没有勾选）。

3. Invert. 已勾选的选择框都被取消，反之亦然。

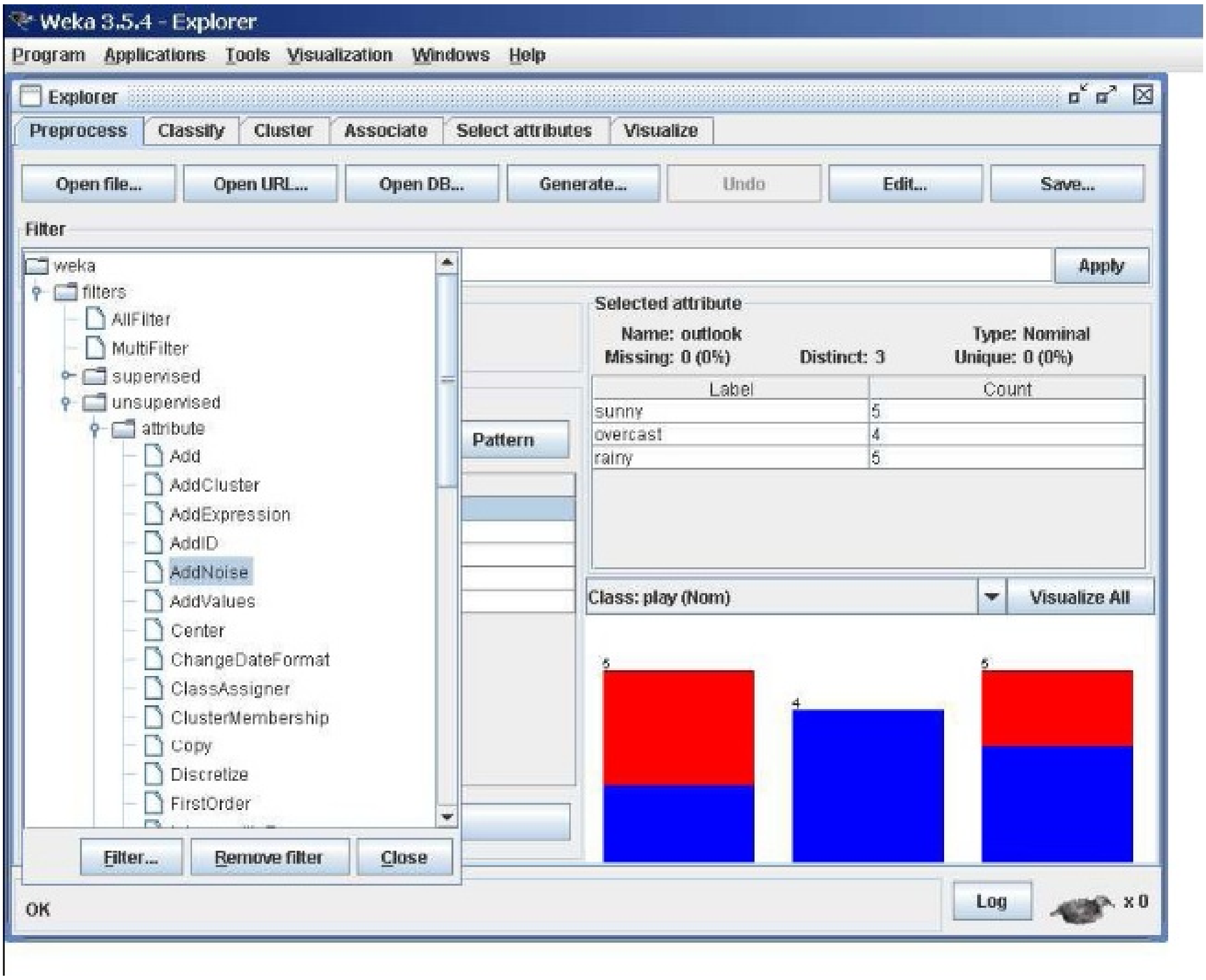
4. Pattern. 让用户基于 Perl 5 正则表达式来选择属性。例如，用 *_id 选择所有名 称以 _id 结束的属性。

选中了想要的属性后，可通过点击属性列表下的 Remove 按钮删除他们。注意可通 过点击位于 Preprocess 面板的右上角的 Edit 按钮旁的 Undo 按钮来取消操作。

3.4 使用筛选器

在预处理阶段，可以定义筛选器来以各种方式对数据进行变换。Filter 一栏用于对各

种筛选器进行必要的设置。Filter 一栏的左边是一个 Choose 按钮。点击这个按钮就可 选择 WEKA 中的某个筛选器。选定一个筛选器后，它的名字和选项会显示在 Choose 按 钮旁边的文本框中。用鼠标左键点击这个框，将出现一个 GenericObjectEditor（通用对象 编辑器）对话框。用鼠标右键（或Alt+Shift+左键)点击将出现一个菜单，你可从中选择， 要么在 GenericObjectEditor 对话框中显示相关属性，要么将当前的设置字符复制到剪贴 板。



GenericObjectEditor 对话框

GenericObjectEditor 对话框可以用来配置一个筛选器。同样的对话框也用于配置其他 对象，例如分类器（classifier）和 聚类器（clusterers）(见下文)。窗口中的字段反映了可 用的选项。点击它们中间的一个便可改变 filter 的设置。例如，某项设置可能是一串文本 字符，这时将字符串输入相应的文本框中即可。或者它可能会给出一个下拉框，列出可供 选择的几个状

态。也可能是其他一些操作，根据所需的信息而有所区别。如果把将鼠标指 针停留在某个字段上，会出现一个小提示来给出相应选项的信息。而有关该筛选器和它的 选项的更多信息可通过点击 GenericObjectEditor 窗口顶部 About 面板中的 More 按 钮来获得。

除了 More 按钮，某些对象也会在关于栏中显示一些有关其功能的简短描述。点击 More 按钮，会出现一个窗口来描述了不同的选项分别起什么作用。有的还另外一个 Capabilities 按钮，它能列出该对象可处理的属性和 class 属性的类型。

GenericObjectEditor 对话框的底部有4个按钮。前两个 Open... 和 Save... 允许存储 对该对象的配置，以备将来之用。Cancel 按钮用于直接退出，任何已作出的改变都将被 忽略。当前选择的对象和设置令人满意后，点击 OK 返回到主 Explorer 窗口。

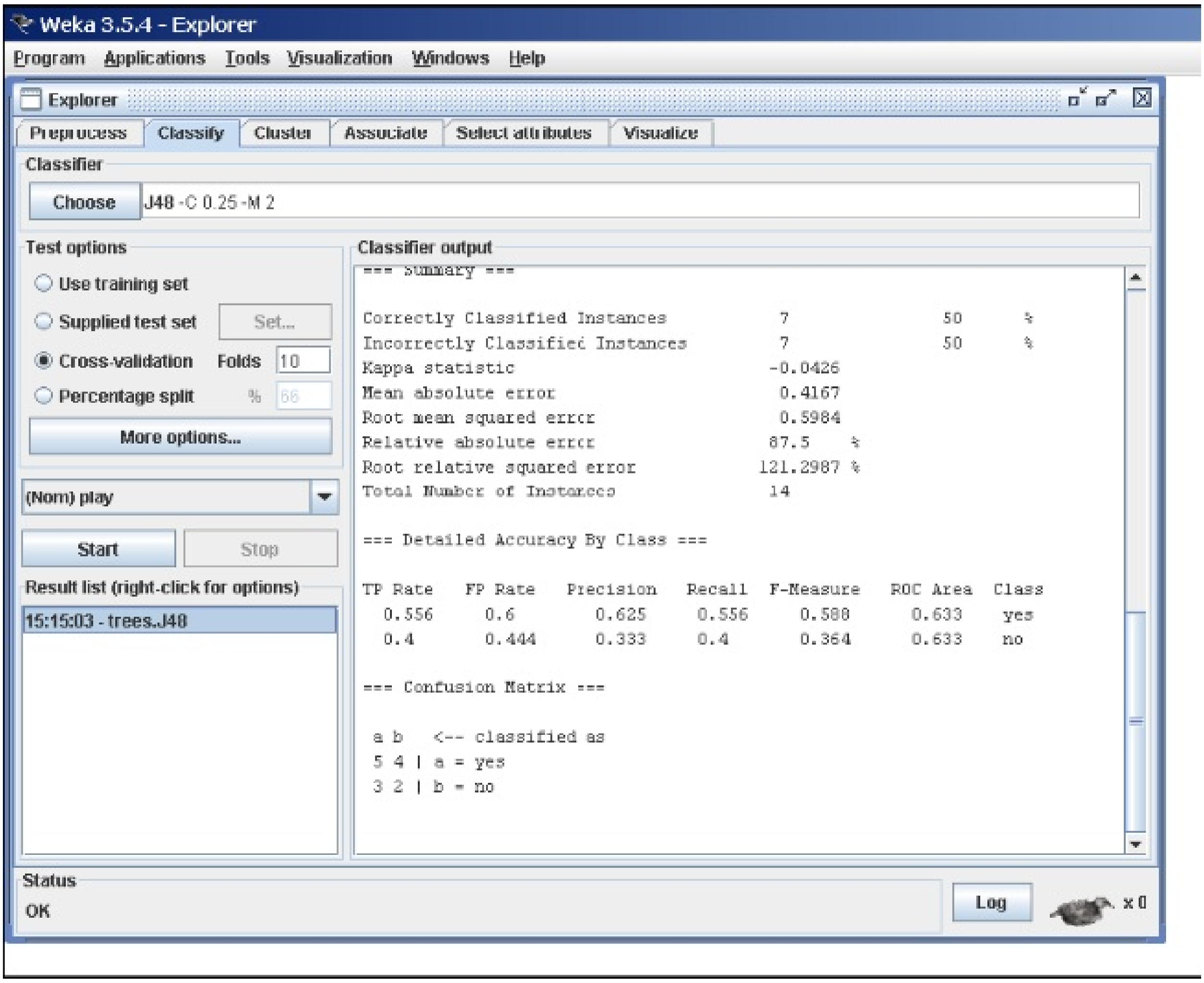
应用筛选器

选择并配置好一个筛选器后，就可通过点击 Preprocess 面板的 Filter 拦右边的 Apply 按钮将之应用于数据集上。然后 Preprocess 面板将显示转换过的数据。可点击 Undo 按钮取消改变。你也可使用 Edit... 按钮在一个数据集编辑器中手动修改你的数 据。最后，点击 Preprocess 面板右上角的 Save... 按钮将用同样的格式保存当前的关系， 以备将来使用。

注意：一些筛选器会依据是否设置了 class 属性来做出不同的动作。（点击直方图上 方那一栏时，会出现一个可供选择的下拉列表。）特别的， `supervised filters` （监督 式筛选器）需要设置一个 class 属性， 而某些`unsupervised attribute filters` （非监督 式属性筛选器）将忽略 class 属性。注意也可以将 Class 设成 None，这时没有设置 class 属性。

> 3 筛选器的英文原文是 filter，与数据库术语中的筛选有关。但是 WEKA 中的 filter 不仅能提供筛选功能， 还涵盖了其他各种数据变换。—译注。

4 分类


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
15:15:03 - trees.J48
Classifier output
Correctly Classified Instances 7 50 %
Incorrectly Classified Instances 7 50 %
Kappa statistic -0.0426
Mean absolute error 0.4167
Root mean squared error 0.5984
Relative absolute error 87.5 %
Root relative squared error 121.2987 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.556 0.6 0.625 0.556 0.588 0.633 yes
0.4 0.444 0.333 0.4 0.364 0.633 no
=== Confusion Matrix ===
a b <-- classified as
5 4 | a = yes
3 2 | b = no
Status
OK
Log
x 0

4.1 选择分类器

在 classify 页面的顶部是 Classifier 栏。这一栏中有一个文本框，给出了分类器的名

称和它的选项。左键点击文本框会打开一个 GenericObjectEditor，可以像配置筛选器那样 配置当前的分类器。右键（或Alt+Shift+左键)点击也可以复制设置字符到剪贴板，或者在 GenericObjectEditor 中显示相关属性。Choose 按钮用来选择 WEKA 中可用的分类器。

4.2 测试选项

应用选定的分类器后得到的结果会根据 Test Option 一栏中的选择来进行测试。共有四种测试模式：

1. Using training set. 根据分类器在用来训练的实例上的预测效果来评价它。

2. Supplied test set. 从文件载入的一组实例，根据分类器在这组实例上的预测效 果来评价它。点击 Set… 按钮将打开一个对话框来选择用来测试的文件。

3. Cross-validation. 使用交叉验证来评价分类器，所用的折数填在 Folds 文本框 中。

4. Percentage split. 从数据集中按一定百分比取出部分数据放在一边作测试用，根 据分类器这些实例上预测效果来评价它。取出的数据量由 % 一栏中的值决定。 注意：不管使用哪种测试方法，得到的模型总是从所有训练数据中构建的。点击 More

options 按钮可以设置更多的测试选项：

1. Output model. 输出基于整个训练集的分类模型，从而模型可以被查看，可视化 等。该选项默认诗选中的。

2. Output per-class stats. 输出每个 class 的准确度/反馈率（precision/recall） 和正确/错误（true/false）的统计量。该选项也是默认选中的

3. Output evaluation measures. 输出熵估计度量。该选项默认没有选中。

4. Output confusion matrix. 输出分类器预测结果的混淆矩阵。该选项默认选中。

5. Store predictions for visualization. 记录分类器的预测结果使得它们能被可 视化表示。

6. Output predictions. 输出测试数据的预测结果。注意在交叉验证时，实例的编 号不代表它在数据集中的位置。

7. Cost-sensitive evaluation. 误差将根据一个价值矩阵来估计。Set… 按钮用来 指定价值矩阵。

8. Random seed for xval / % Split. 指定一个随即种子，当出于评价的目的需要 分割数据时，它用来随机化数据。

> 4 WEKA 中的分类和回归都放入了 classify 页面中，相应的工具都叫做分类器（classifier）。参考4.3节。

4.3 Class 属性

WEKA 中的分类器被设计成经过训练后可以预测一个 class 属性，也就是预测的目标。有的分类器只可用来学习分类型的 class 属性；有的则只可用来学习数值型的 class 属性 (回归问题)；还有的两者都可以学习。

默认的，数据集中的最后一个属性被看作 class 属性。如果想训练一个分类器，让它 预测一个不同的属性，点击 Test options 栏下方的那一栏，会出现一个属性的下拉列表 以供选择。

4.4 训练分类器

分类器，测试选项和 class 属性都设置好后，点击 Start 按钮就可以开始学习过程。

分类器忙于训练时，下方的小鸟会动来动去。可以通过点击 Stop 按钮，在任意时刻停止 训练过程。

训练完成后，会发生几件事。右边的 Classifier output 区域会被填充一些文本，描述训 练和测试的结果。在 Result list 栏中会出现一个新的条目。接下来我们会观察这个结果 列表，但我们先来研究输出的文本。

4.5 分类器输出文本

Classifier output 区域的文本有一个滚动条以便浏览结果。按住 Alt 和 Shift 键，在这个区域点击鼠标左键，会出现一个对话框，让你用各种格式（目前可用 JPEG 和 EPS） 保存输出的结果。当然，可以通过放大 Explorer 窗口来获得更大的显示区域。输出结果 可分为几个部分：

1. Run information. 给出了学习算法各选项的一个列表。包括了学习过程中涉及 到的关系名称，属性，实例和测试模式。

2. Classifier model (full training set). 用文本表示的基于整个训练集的分类模 型。

所选测试模式的结果可以分解为以下几个部分：

3. Summary. 一列统计量，描述了在指定测试模式下，分类器预测 class 属性的准 确程度。

4. Detailed Accuracy By Class. 更详细地给出了关于每一类的预测准确度的描 述。

5. Confusion Matrix. 给出了预测结果中每个类的实例数。其中矩阵的行是实际的 类，矩阵的列是预测得到的类，矩阵元素就是相应测试样本的个数。

4.6 结果列表

在训练了若干分类器之后，结果列表中也就包含了若干个条目。左键点击这些条目可以在生成的结果之间进行切换浏览。右键点击某个条目则会弹出一个菜单，包括如下的选 项：

1. View in main window. 在主窗口中显示输出该结果（就象左击该条目一样）。

2. View in separate window. 打开一个独立的新窗口来显示结果。

3. Save result buffer. 弹出一个对话框，使得输出结果的文本可以保存成一个文本 文件。

4. Load model. 从一个二进制文件中载入以前训练得到的模型对象。

5. Save model. 把模型对象保存到一个二进制文件中。对象是以 Java `序列化` 的形式保存的。

6. Re-evaluate model on current test set. 通过 Supplied test set 选项下的 Set 按钮指定一个数据集，已建立的分类模型将在这个数据集上测试它的表现。

7. Visualize classifier errors. 弹出一个可视化窗口，把分类结果做成一个散点图。 其中正确分类的结果用叉表示，分错的结果用方框表示。

8. Visualize tree or Visualize graph. 如果可能的话，把分类模型的结构用图形来 表示（例如决策树（decision tree）和贝叶斯网络（Bayesian network）模型）。 图形可视化选项只有在贝叶斯网络模型建好之后才会出现。在浏览决策树图形时， 可以在空白处右击鼠标弹出一个菜单，也可以拖动鼠标来拖动决策树，还可以在 节点上单击鼠标查看它对应的训练实例。Ctrl键+左键点击会缩小图形，Shift键+ 拖曳会得到一个方框并放大其中的图形。这个图形可视化工具本身应该能够解释 它的作用。

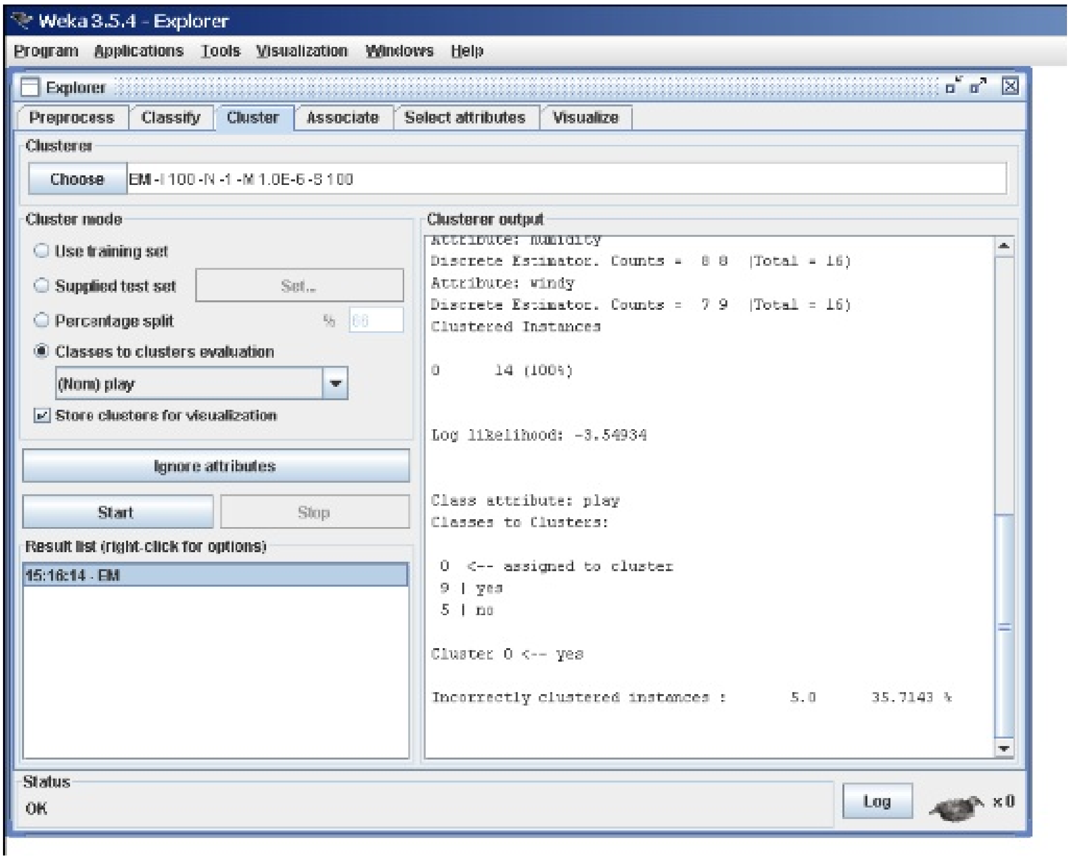
9. Visualize margin curve. 创建一个散点图来显示预测边际值。这个边际值的定 义为：预测为真实值的概率与预测为真实值之外其它某类的最高概率之差。例如， 提升式（boosting）算法可以通过增加训练数据上的边际值来使得它在测试数据上 表现得更好。

10. Visualize threshold curve. 生成一个散点图，以演示预测时改变各类之间的阀 值后取得的平衡。例如说，默认的阀值是0.5，那么一个实例要预测成为 `positive` ， 它是 `positive` 的预测概率必须大于0.5。这个曲线可以用来在 ROC 曲线分析中 演示准确度/反馈率之间的平衡（正确的 positive 率对错误的 positive 率），也 可用于其它类型的曲线。

11. Visualize cost curve. 生成一个散点图，如 [1] 中所描述的那样，给出期望价 值（expected cost）的一个显式表达。

在特定的情况下某些选项不适用时，它们会变成灰色。

5 聚类


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Clusterer
Choose
EM -I 100 -N -1 -M 1.0E-6 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Nom) play
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
15:16:14 - EM
Clusterer output
Attribute: humidity
Discrete Estimator. Counts = 8 8 (Total = 16)
Attribute: windy
Discrete Estimator. Counts = 7 9 (Total = 16)
Clustered Instances
0 14 (100%)
Log likelihood: -3.54934
Class attribute: play
Classes to Clusters:
0 <-- assigned to cluster
9 | yes
5 | no
Cluster 0 <-- yes
Incorrectly clustered instances : 5.0 35.7143 %
Status
OK
Log
x 0

5.1 选择聚类器（Clusterer）

现在我们应该熟悉选择和配置对象的过程了。点击列在窗口顶部的 Clusterer 栏中的聚类算法，将弹出一个用来选择新聚类算法的 GenericObjectEditor 对话框。

5.2 聚类模式

Cluster Mode 一栏用来决定依据什么来聚类以及如何评价聚类的结果。前三个选项和分类的情形是一样的：Use training set、 Supplied test set 和 Percentage split （见4.1节）——区别在于现在的数据是要聚集到某个类中，而不是预测为某个指定的类 别。第四个模式，Classes to clusters evaluation，是要比较所得到的聚类与在数据中 预先给出的类别吻合得怎样。和 Classify 面板一样，下方的下拉框是用来选择作为类别 的属性的。

在 Cluster mode 之外，有一个 Store clusters for visualization 的勾选框，该框决定 了在训练完算法后可否对数据进行可视化。对于非常大的数据集，内存可能成为瓶颈时， 不勾选这一栏应该会有所帮助。

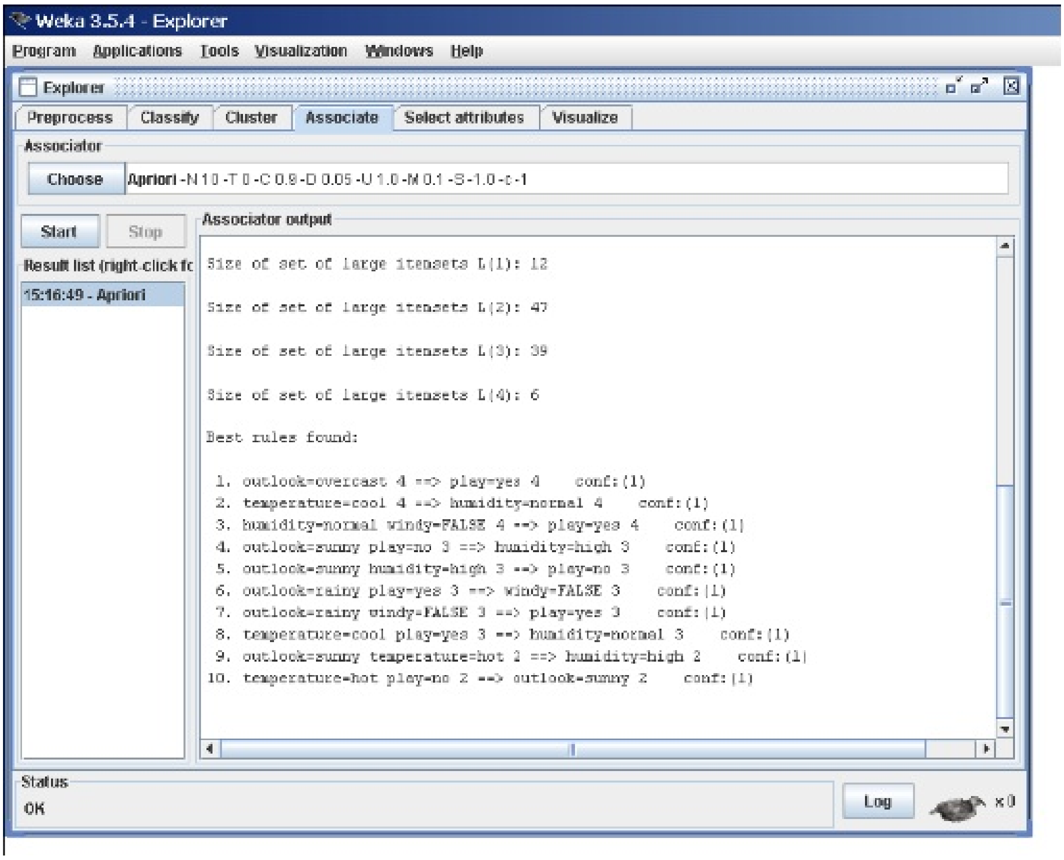
5.3 忽略属性

在对一个数据集聚类时，经常遇到某些属性应该被忽略的情况。Ignore attributes 可以弹出一个小窗口，选择哪些是需要忽略的属性。点击窗口中单个属性将使它高亮显示， 按住 SHIFT 键可以连续的选择一串属性，按住 CTRL 键可以决定各个属性被选与否。点 击 Cancel 按钮取消所作的选择。点击 Select 按钮决定接受所作的选择。下一次聚类 算法运行时，被选的属性将被忽略。

5.4 学习聚类

Cluster 面板就像Classify面板那样，有一个 Start/Stop 按钮，一个结果文本的区域和一个结果列表。它们的用法都和分类时的一样。右键点击结果列表中的一个条目将弹 出一个相似的菜单，只是它仅显示两个可视化选项：Visualize cluster assignments 和 Visualize tree。后者在它不可用时会变灰。

6 关联规则


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Result list (right-click fo
15:16:49 - Apriori
Associator output
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 47
Size of set of large itemsets L(3): 39
Size of set of large itemsets L(4): 6
Best rules found:
1. outlook=overcast 4 ==> play=yes 4 conf:(1)
2. temperature=cool 4 ==> humidity=normal 4 conf:(1)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 conf:(1)
4. outlook=sunny play=no 3 ==> humidity=high 3 conf:(1)
5. outlook=sunny humidity=high 3 ==> play=no 3 conf:(1)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 conf:(1)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 conf:(1)
8. temperature=cool play=yes 3 ==> humidity=normal 3 conf:(1)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2 conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2 conf:(1)
Status
OK
Log
x 0

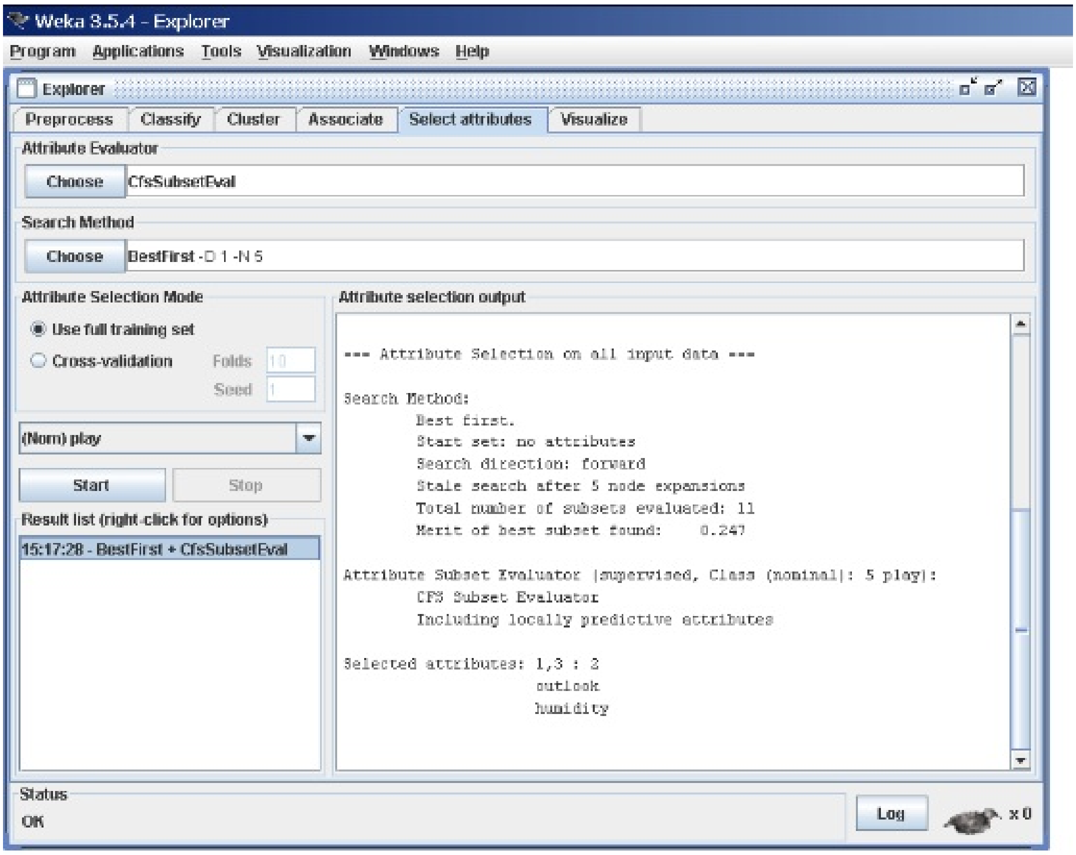
6.1 设定

这个面板包含了学习关联规则的方案。这里的学习器也可以跟其它面板的聚类器，筛选器和分类器一样选择和配置。

6.2 学习关联规则

为关联规则学习器设置好合适的参数后，点击 Start 按钮。完成后右键点击结果列表中的条目可以查看或保存结果。

7 属性选择


Weka 3.5.4 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Attribute Evaluator
Choose CfsSubsetEval
Search Method
Choose BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation Folds 10
Seed 1
(Nom) play
Start Stop
Result list (right-click for options)
15:17:28 - BestFirst + CfsSubsetEval
Attribute selection output
=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 0.247
Attribute Subset Evaluator (supervised, Class (nominal): 5 play):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 1,3 : 2
outlook
humidity
Status
OK
Log
x 0

7.1 搜索与评估

属性选择是说搜索数据集中全部属性的所有可能组合，找出预测效果最好的那一组属性。为实现这一目标，必须设定两个东西：属性评估器（evaluator）和搜索策略。评估器 决定了怎样给一组属性安排一个表示它们好坏的值。搜索策略决定了要怎样进行搜索。

7.2 选项

Attribute Selection Mode 一栏有两个选项。

1. Use full training set. 使用训练数据的全体好决定一组属性的好坏。

2. Cross-validation. 一组属性的好坏通过一个交叉验证过程来决定。Fold 和

Seed 分别给出了交叉验证的折数和打乱数据时的随机种子。 和 Classify 部分（4.1节）一样，有一个下拉框来指定 class 属性。

7.3 执行选择

点击 Start 按钮开始执行属性选择过程。它完成后，结果会输出到结果区域中，同时结果列表中会增加一个条目。在结果列表上右击，会给出若干选项。其中前面三个（View in main window，View in separate window 和 Save result buffe）和分类面板中 是一样的。还可以可视化精简过的数据集（Visualize reduced data），或者，如果使用 过主成分分析那样的属性变换工具，则能可视化变换过的数据集（Visualize transformed data）。精简过/变换过的数据能够通过 Save reduced data... 或 Save transformed data... 选项来保存。

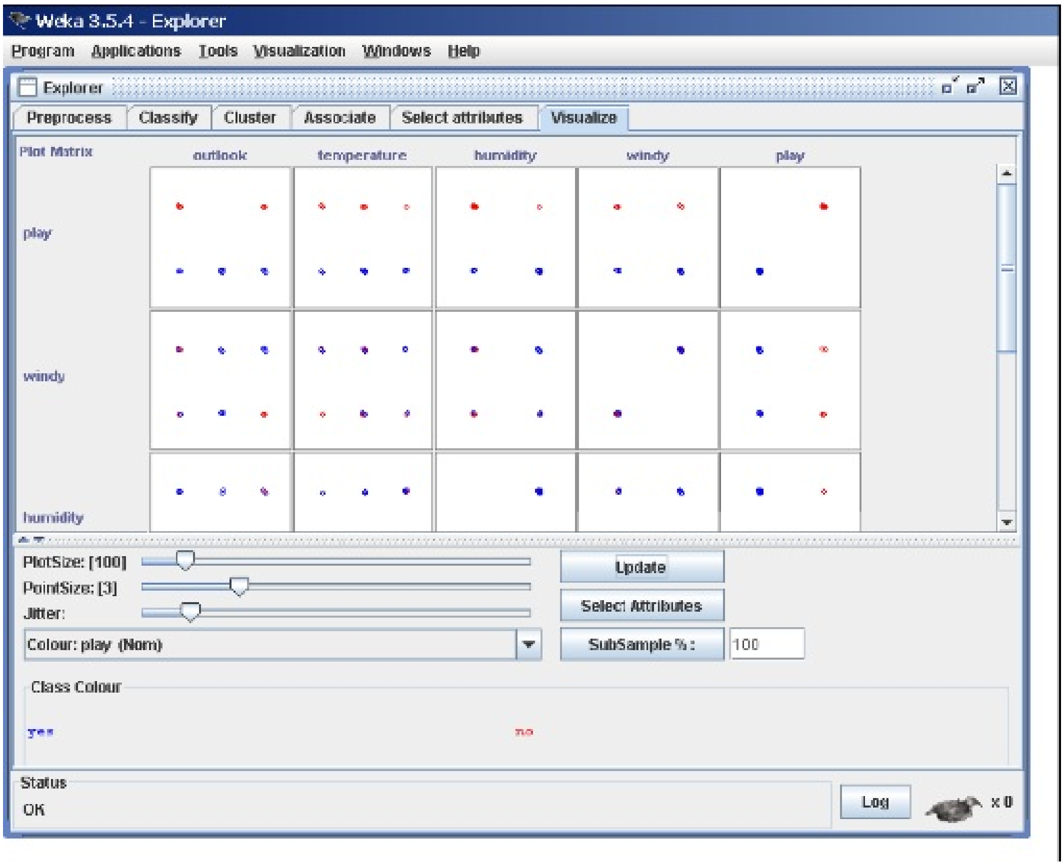
如果想在精 简 / 变换训 练集的同时 进行测试， 而又不使用 在分类器面 板中的 AttributeSelectedClassifier，那么最好在命令行或者 SimpleCLI 中使用批量模式（ `-b` ） 的 AttributeSelection 筛选器（这是一个 supervised attribute filter）。这一批量模式允许 指定额外的输入和输出文件对（选项 –r 和 -s），处理它们的筛选器的设置是由训练文件（由 –i 和 –o 选项给出）决定的。 下面是 Unix/Linux bash 中的一个例子：

```
java weka.filters.supervised.attribute.AttributeSelection \
        -E "weka.attributeSelection.CfsSubsetEval " \
        -S "weka.attributeSelection.BestFirst -D 1 -N 5" \
        -b \
        -i <input1.arff> \
        -o <output1.arff> \
        -r <input2.arff> \
        -s <output2.arff>
```

注意：

- 每一样末尾的反斜线告诉 bash 命令还没有结束。使用 SimpleCLI 时必须把命令 写在同一行而不能使用反斜线。
- 这里假设 WEKA 已经在 CLASSPATH 中了，否则还要加上 -classpath 选项。
- 整个筛选器的设置会在日志中输出，就像运行正规的属性选择时的设置一样。

8 可视化



WEKA 的可视化页面可以对当前的关系作二维散点图式的可视化浏览。

8.1 散点图矩阵

选择了 Visualize 面板后，会为所有的属性给出一个散点图矩阵，它们会根据所选的 class 属性来着色。在这里可以改变每个二维散点图的大小，改变各点的大小，以及随机 地抖动（jitter）数据（使得被隐藏的点显示出来）。也可以改变用来着色的属性，可以只 选择一组属性的子集放在散点图矩阵中，还可以取出数据的一个子样本。注意这些改变只 有在点击了 Update 了按钮之后才会生效。

8.2 选择单独的二维散点图

在散点图矩阵的一个元素上点击后，会弹出一个单独的窗口对所选的散点图进行可视化。（前面我们描述了如何在单独的窗口中对某个特定的结果进行可视化——例如分类误 差——这里用了相同的可视化控件。）

数据点散布在窗口的主要区域里。上方是两个下拉框选择用来选择打点的坐标轴。左 边是用作 x 轴的属性；右边是用作 y 轴的属性。

在 x 轴选择器旁边是一个下拉框用来选择着色的方案。它可以根据所选的属性给点着 色。在打点区域的下方，有图例来说明每种颜色代表的是什么值。如果这些值是离散的， 可以通过点击它们所弹出的新窗口来修改颜色。

打点区域的右边有一些水平横条。每一条代表着一个属性，其中的点代表了属性值的 分布。这些点随机的在竖直方向散开，使得点的密集程度能被看出来。在这些横条上点击 可以改变主图所用的坐标轴。左键点击改变 x 轴的属性；右键点击改变 y 轴的。横条旁边的 X 和 Y 代表了当前的轴用的那个属性（ B 则说明 x 轴和 y 轴都是它）。 属性横条的上方是一个标着 Jitter 的游标。它能随机地使得散点图中各点的位置发生偏移，也就是抖动。把它拖动到右边可以增加抖动的幅度，这对识别点的密集程度很有用。 如果不使用这样的抖动，几万个点放在一起和单独的一个点看起来会没有区别。

8.3 选择实例

很多时候利用可视化工具选出一个数据的子集是有帮助的。（例如在 classify 面板的 UserClassifier（自定义分类器），可以通过交互式的选取实例来构建一个分类器。）

在 y 轴选择按钮的下方是一个下拉按钮，它决定选取实例的方法。可以通过以下四种 方式选取数据点：

1. Select Instance. 点击各数据点会打开一个窗口列出它的属性值，如果点击处的 点超过一个，则更多组的属性值也会列出来。

2. Rectangle. 通过拖动创建一个矩形，选取其中的点。

3. Polygon. 创建一个形式自由的多边形并选取其中的点。左键点击添加多边形的顶 点，右键点击完成顶点设置。起始点和最终点会自动连接起来因此多边形总是闭 合的。

4. Polyline. 可以创建一条折线把它两边的点区分开。左键添加折线顶点，右键结束 设置。折线总是打开的（与闭合的多边形相反）。

使用 Rectangle，Polygon 或 Polyline 选取了散点图的一个区域后，该区域会变 成灰色。这时点击 Submit 按钮会移除落在灰色区域之外的所有实例。点击 Clear 按钮 会清除所选区域而不对图形产生任何影响。

如果所有的点都被从图中移除，则 Submit 按钮会变成 Reset 按钮。这个按钮能使 前面所做的移除都被取消，图形回到所有点都在的初始状态。最后，点击 Save 按钮可 把当前能看到的实例保存到一个新的 ARFF 文件中。

参考文献

[1] Drummond, C. and Holte, R. (2000) Explicitly representing expected cost: An alternative to ROC representation. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Publishers, San Mateo, CA.

[2] Witten, I.H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques. 2nd edition* Morgan Kaufmann, San Francisco.

[3] *WekaWiki* – `http://weka.sourceforge.net/wiki/`

[4] *WekaDoc* – `http://weka.sourceforge.net/wekadoc/`

[5] Ensemble Selection on *WekaDoc* –
`http://weka.sourceforge.net/wekadoc/index.php/en:Ensemble_Selection`